Homework 6: Outlier Detection

About

Due

Monday 3/11/19, 11:59 PM CST

Goal

This homework focuses on implementing linear regression and using Cook's distance, leverage and standardized residuals to dectect outliers by R. **NOTE: you are required to use R in this homework. Otherwise you will receieve ZERO mark.**

R Tutorials

- Quick R Tutorial (https://www.statmethods.net/r-tutorial/index.html)
- Learn R in Several Minutes (https://learnxinyminutes.com/docs/r/)
- Outliers Diagnostics by R (https://data.library.virginia.edu/diagnostic-plots/)

Problems

At https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.data (https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.data), you will find the famous Boston Housing dataset. This consists of 506 data items. You will find the explanation of the dataset at https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names (https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names). Each is 13 measurements, and a house price. The data was collected by Harrison, D. and Rubinfeld, D.L in the 1970s (a date which explains the very low house prices). The dataset has been widely used in regression exercises, but seems to be waning in popularity. At least one of the independent variables measures the fraction of population nearby that is ''Black'' (their word, not mine). This variable appears to have had a significant effect on house prices then (and, sadly, may still now).

1. (**50 points**) Regress house price (variable 14) against all others, and use leverage, Cook's distance, and standardized residuals to find possible outliers. Use `plot(your_linear_regression_model)` to produce a diagnostic plot that allows you to identify possible outliers (points with high residual or high leverage or high influence).

Give the indices of possible outliers and **explain** why you think they are outliers. The reason being an outlier may not be same, so please be sure you understand the plot well. (You can read R tutorial provided above if you have any questions).

2. (**30 points**) Remove all points you suspect as outliers, and compute a new regression. Reproduce a diagnostic plot that allows you to identify possible outliers. We do not require explanations this time.

3. (**10 points**) Apply a Box-Cox transformation (use `boxcox` command) to the dependent variable, what is the best value of the parameter?

4. (**10 points**) Now transform the dependent variable, build a linear regression, and check the standardized residuals. If they look acceptable, produce a plot of fitted house price against true house price.

Submission

Submission will be through gradescope (https://www.gradescope.com/). Your submission should be a PDF with the following pages.

1. (**0 points**) **Page 1**: code for regression and resulting model.

2. (**50 points**) **Page 2**: a screenshot of your diagnostic plot and a few sentences of your explanation.

3. (**20 points**) **Page 3**: a screenshot of your new diagnostic plot.

4. (**10 points**) **Page 4**: a screenshot of your code for subproblem 2.

5. (**10 points**) **Page 5**: a screenshot of Box-Cox transformation plot and the best value you chose.

6. (**10 points**) **Page 6**: result of the standardized residuals of the regression after Box-Cox transformation and a plot of fitted house price against true house price.

7. (**0 points**) **Page 7**: code for subproblems 3 and 4.